Age Estimation Using Expectation of Label Distribution Learning

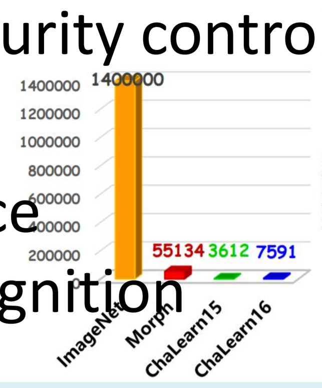
Bin-Bin Gao[1], Hong-Yu Zhou[1], Jianxin Wu[1], Xin Geng[2]
[1]National Key Laboratory for Novel Software Technology, Nanjing University
[2]Southeast University

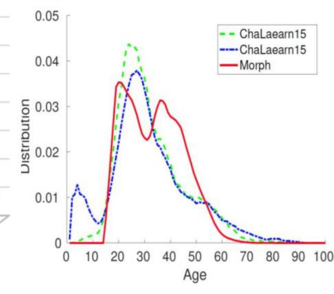
Learning And Mining from DatA
http://lamda.nju.edu.cn

1. Background

- Face information
  Identity, Emotion, Ethnicity, Gender, Attractiveness and Age etc.

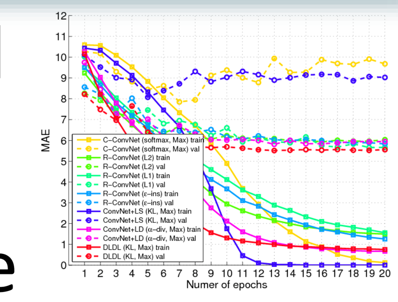
- Facial age estimation has many applications
  Law enforcement, Security control and Recommendations.
- Challenges
  Insufficiency, Imbalance and Fine-grained Recognition





2. Motivation

- Regression and classification may lead to an **unstable** training procedure.

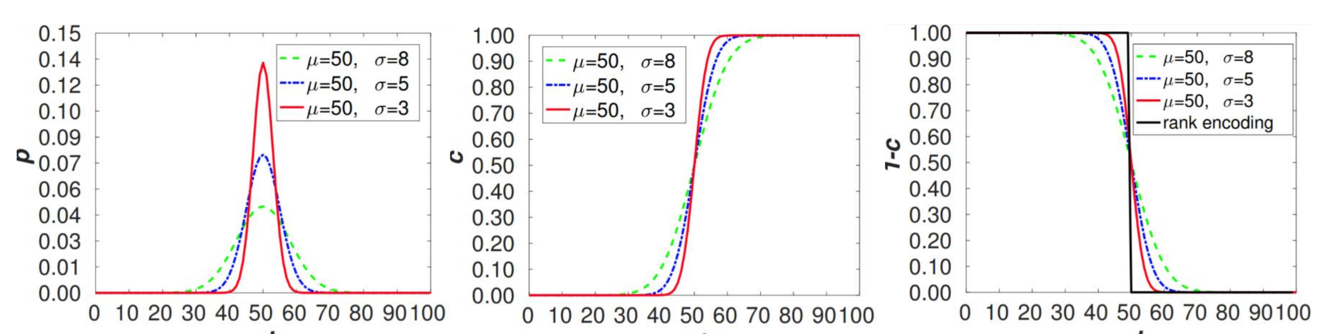


- There is an **inconsistency** between the training objectives and evaluation metric in DLDL and Ranking.

- Almost all state-of-the-arts have **huge computational and storage cost**.

3. Proposed Method: DLDL-v2

- Ranking is Learning Label Distribution



$$p^{ld} = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(l-y)^2}{2\sigma^2}\right) \quad c = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{l-y}{\sigma\sqrt{2}}\right)\right] \quad p^{rank} = [\underbrace{1,\ldots,1}_{\lfloor y \rfloor}, \underbrace{0,\ldots,0}_{\lfloor K-1-y \rfloor}]$$

$$\mathbf{c} = \mathbf{T}\boldsymbol{p}^{ld} \qquad \boldsymbol{p}^{rank} \approx \mathbf{1} - \mathbf{c}$$

$$\boldsymbol{p}^{rank} \approx \mathbf{1} - \mathbf{T}\boldsymbol{p}^{ld}$$

There is a linear relationship between Ranking encoding and label distribution (LD).

✓ LD represents more meaningful information.
✓ LD learning is more efficient.

- Joint Learning for Age Estimation

✓ Distribution Learning Module

- Linear transformation

$$\boldsymbol{x} = \mathbf{W}^T \boxed{\boldsymbol{f}} + \boldsymbol{b} \quad \text{CNN feature}$$

- Label distribution

$$\hat{p}_k = \boxed{\frac{\exp(x_k)}{\sum_t \exp(x_t)}} \quad \text{Softmax}$$

- Loss: KL-Div

$$L_{ld} = \sum_k p_k \ln \frac{\boxed{p_k}}{\boxed{\hat{p}_k}} \quad \text{Label Dis} \quad \text{Pred Dis}$$

✓ Expectation Regression Module

- Expectation layer

$$\hat{y} = \sum_k \hat{p}_k \boxed{l_k} \quad \text{Label Set}$$

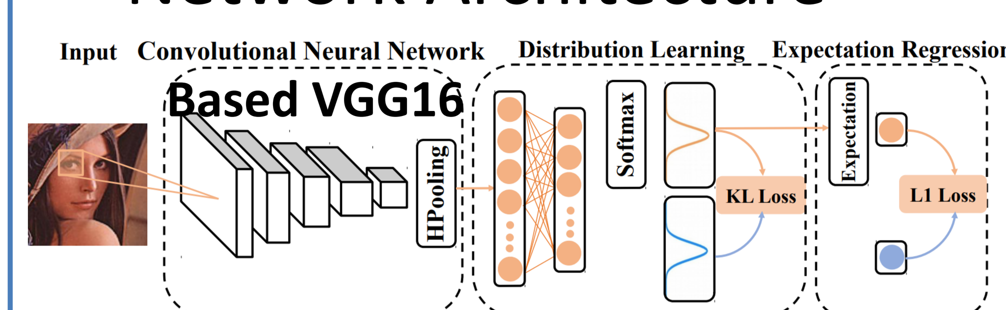
- Loss: ℓ_1

$$L_{er} = |\hat{y} - y|$$

✓ Learning (SGD algorithm)

$$L = L_{ld} + \lambda L_{er}$$

- Network Architecture



- Remove all FCs instead of a hybrid-pooling layer.
- Reduce the number of the filters in each Conv layer.
- Add a BN layer after each Conv layer.
- Add DL and ER modules.

4. Experiments

- Three benchmark age datasets.
  ChaLearn15, ChaLearn16 and Morph.
- Comparisons with State-of-the-Arts.

Table 1: Comparisons on apparent and real age estimation.

Methods	External Data	ChaLearn15 MAE	ChaLearn15 ϵ-error	ChaLearn16 MAE	ChaLearn16 ϵ-error	Morph MAE
Human [Han *et al.*, 2015]	×	-	0.34	-	-	6.30
OR-CNN [Niu *et al.*, 2016]	×	-	-	-	-	3.34
DEX [Rothe *et al.*, 2018]	×	5.369	0.456	-	-	3.25
DEX [Rothe *et al.*, 2018]	√	3.252	0.282	-	-	2.68
DLDL [Gao *et al.*, 2017]	×	3.51	0.31	-	-	2.42[1]
Ranking [Chen *et al.*, 2017]	×	-	-	-	-	2.96
LDAE [Antipov *et al.*, 2017]	√	-	-	-	0.241[2]	2.35
DLDL-v2 (TinyAgeNet)	×	3.427	0.301	3.765	0.291	2.291
DLDL-v2 (ThinAgeNet)	×	**3.135**	**0.272**	**3.452**	**0.267**	**1.969**

[1]Used 90% of Morph images for training and 10% for evaluation;
[2]Used multi-model ensemble;

Low Error

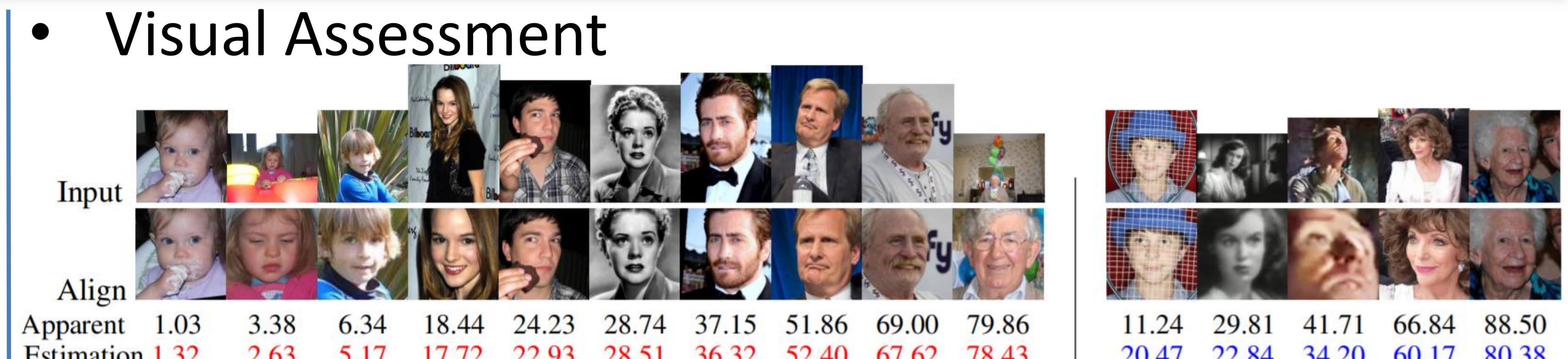
Table 2: Comparisons of model parameters and forward times. (32 images on M40)

Methods	#Param(M)	#Time(ms)
DEX [Rothe *et al.*, 2018]	134.6	133.30
DLDL [Gao *et al.*, 2017]	134.6	133.30
LDAE [Antipov *et al.*, 2017]	1480.6	1446.30
DLDL-v2 (TinyAgeNet)	0.9	24.26
DLDL-v2 (ThinAgeNet)	3.7	51.05

Interpretability

High Efficiency

- Visual Assessment



- Ablation Study

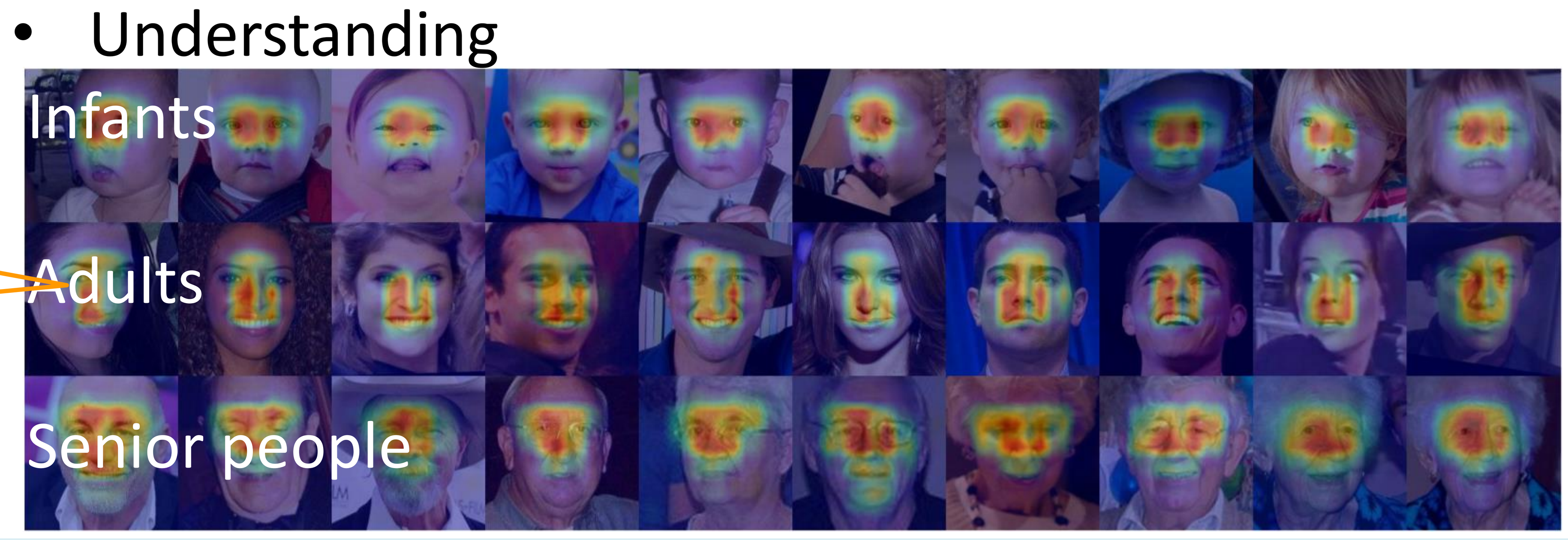
Comparisons with Baselines

Methods	Aug	Pool	ChaLearn15 MAE	ChaLearn15 ϵ-error	ChaLearn16 MAE	ChaLearn16 ϵ-error	Morph MAE
DLDL-v2	×	HP	3.399	0.303	3.717	0.290	2.346
	√	GAP	3.210	0.282	3.539	0.274	2.039
	√	HP	**3.135**	**0.272**	**3.452**	**0.267**	**1.969**
MR (ℓ_2)	√	HP	3.665	0.337	3.696	0.294	2.282
MR (ℓ_1)	√	HP	3.655	0.334	3.722	0.301	2.347
DEX	√	HP	3.558	0.306	4.163	0.332	2.311
Ranking	√	HP	3.365	0.298	3.645	0.290	2.164
ER (ℓ_1)	√	HP	3.287	0.291	3.641	0.282	2.214
DLDL	√	HP	3.228	0.285	3.509	0.272	2.132

Sensitivity to Hyper-parameters

λ	Δl (K)	ChaLearn15 MAE	ChaLearn15 ϵ-error	ChaLearn16 MAE	ChaLearn16 ϵ-error	Morph MAE
0.01	1 (101)	3.223	0.282	3.493	0.270	**1.960**
0.10	1 (101)	3.188	0.278	3.455	0.268	1.972
1.00	1 (101)	**3.135**	**0.272**	**3.452**	0.267	1.969
10.00	1 (101)	3.144	0.273	3.487	0.270	1.977
1.00	4 (26)	3.182	0.276	3.473	0.270	1.963
1.00	2 (51)	3.184	0.274	3.484	0.271	1.963
1.00	0.50 (201)	3.184	0.278	3.484	0.269	1.992
1.00	0.25 (401)	3.167	0.274	3.459	**0.265**	2.028

- Understanding



5. Conclusion

✓ We provide the first analysis and show that the ranking method is in fact learning label distribution implicitly. This result thus unifies existing state-of-the-art facial age estimation methods into the DLDL framework.
✓ We propose an end-to-end learning framework which jointly learns age distribution and regresses single-value age in both feature learning and classifier learning.
✓ We create new state-of-the-art results on facial age estimation tasks using single and small model without external age labeled data or multi-model ensemble.

Project